Morphology

David Yarowsky

9/8/2020

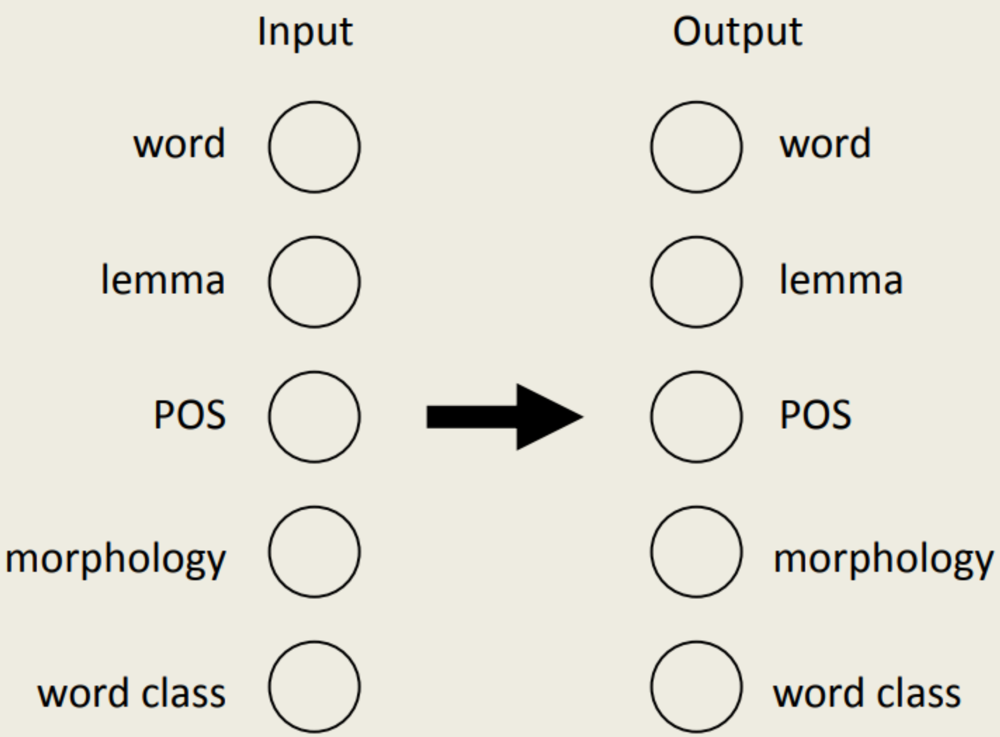
Acknowledgements and thanks to:

- Chris Quirk
- Marta Costa-jussa
- Richard DeArmond
- Eleni Miltsakaki
- Antske Fokkens
- Penny Eckert
- Ivan Sag
- Eleni Miltsakaki
- Adam Szczegielniak
- Jeff Conn
- Dan Jurafsky
- Jason Eisner

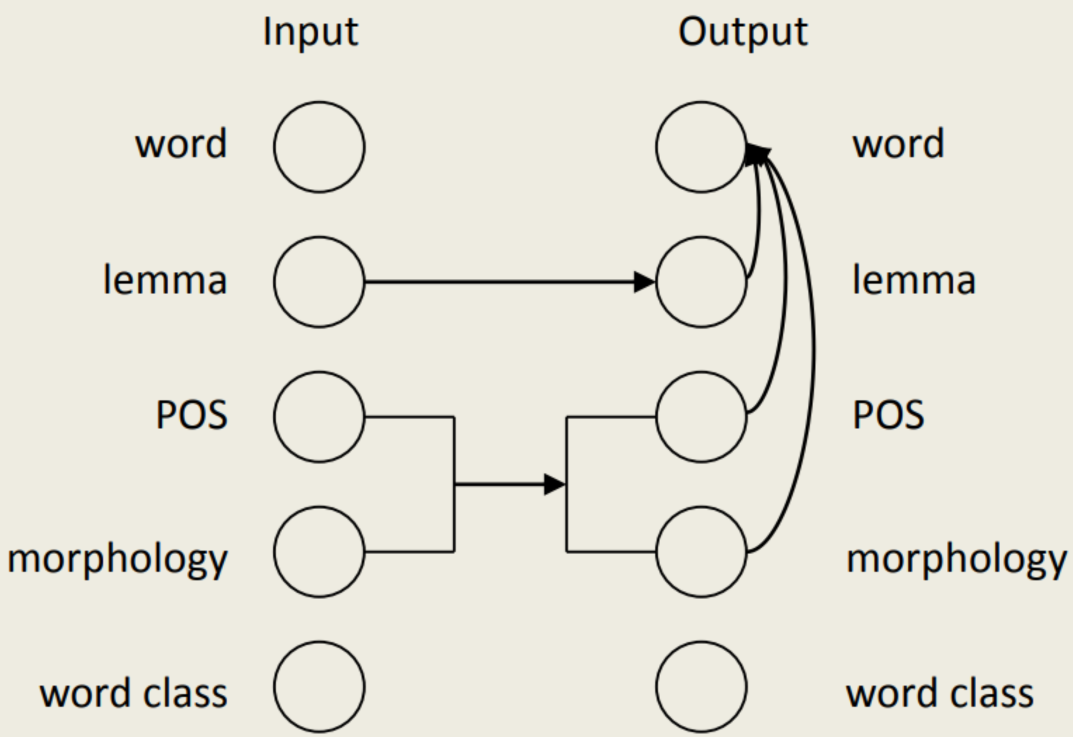
Factored translation models

(and factored language models)

Factored Representation



Factored Model: transfer and generation



Morphology: The Study of Word Structure

How words are put together out of smaller pieces that linguists call **morphemes**, the **minimal** units of linguistic form and meaning.

Morphological Analysis

	morphemes	or	semantic features
dogs	=> dog+s	or	dog+PL
walking	=> walk+ing	or	walk+PRS;PTCP

running => runn+ing?

run+ing & n->nn (gemination)

dancing => danc+ing?

dance+ing & e->NULL (elision)

Morphological Generation

morphemes or semantic features

dog+s => dogs or dog+PL => dogs

walk+ing => walking or walk+PRS;PTCP => walking

run+ing => running or run+PRS;PTCP => runnning

& n->nn (gemination)

dance+ing => dancing or dance+PRS;PTCP => dancing

& e->NULL (elision)

Inflectional Morphology

	morphemes	or semantic features
dogs	=> dog+s	or dog+PL
walking	=> walk+ing	or walk+PRS;PTCP

<= regular grammatical feature extension of same core word meaning

("I am walking" and "I walked" differ only by tense)

inflectional paradigm:

VERB	+PRS;3SG	+PRS;PTCP	+PST;PFV	+PST;PTCP
	(+s)	(+ing)	(+ed)	(+en/+ed)
walk eat	walks eats	walking eating	walked ate	walked eaten

<= canonical affixes

Inflectional Morphology

	morphemes	or semantic features
dogs	=> dog+s	or dog+PL
walking	=> walk+ing	or walk+PRS;PTCP

<= regular grammatical feature extension of same core word meaning

("I am walking" and "I walked" differ only by tense)

inflectional paradigm:

VERB	+PRS;3SG	+PRS;PTCP	+PST;PFV	+PST;PTCP
	(+s)	(+ing)	(+ed)	(+en/+ed)
walk	walks	walking	walked	walked
eat	eats	eating	ate	eaten

<= canonical affixes

Derivational Morphology (new concept formation)

employer => employ+er or employ+V:N(Agent)

employment => employ+ment or employ+V:N(Result/ActOf)

"employ" = An ACTION (verb)

"employer" = A PERSON (noun)

a "dogfight" is not a "dog"

employable => un+[employ+able] (not able to be employed)

[un+employ]+able (able to be not employed?)

<= is "to unemploy" a verb?

Morphological Segmentation

- pre+pose
- pre+pos+ition
- pre+pos+ition+al
- pre+pos+ition+al+ize
- pre+pos+ition+al+iz+ation
- pre+pos+ition+al+iz+ation+free
- Pseudopseudohypoparathyroidism

Morphological Parse

- pre+pose
- [pre+pos]+ition
- [[[pre+pos]+ition]+al]
- [[[[pre+pos]+ition]+al]+ize]
- [[[[[pre+pos]+ition]+al]+iz]+ation]
- [[[[[[pre+pos]+ition]+al]+iz]+ation]+free]
- [[[Pseudo+[pseudo+[hypo+[para+[thyr+oid]]]]]] +ism]

All languages have phonology, syntax and semantics...

- [t] vs. [t^h] vs. [d]
- English is SVO; Irish is VSO; Japanese is SOV.
- [ku]
  - pigeon sound, government takeover, ...
  - blow, punch, neck, ...
  - cow, ...
  - bank, library, ...
- But..... Do all languages have morphology?

Mandarin

(Sino-Tibetan - 845,500,000 speakers)

na^{4}er^5 you^3 gou^3
there have dog
'there's a dog (or dogs) there.'
na^{4}er^5 you^3 ji^3 zhi^1 gou^3
there have several CLASSIFIER dog
'there are dogs there.'

These languages are called **Analytic** (or **Isolating**).

Synthetic Languages

Have affixes (or other **bound** elements) that get attached to other morphemes to build words. There are three kinds:

- Agglutinating Languages
- Fusional Languages
- Polysynthetic Languages

Agglutinating Languages

- The morphemes are put together “loosely”.
- The segmentation of individual morphemes is straightforward, e.g. **Hungarian** (Uralic - 12,500,000 speakers):

[haːz-unk] house-our

[haːz-ɔd] house-your

[haːz-unk-bɔn] house-our-in

[haːz-od-bɔn] house-your-in

More Hungarian

- [taːrʃ] (‘companion’)
- [taːrʃ + ɔs (‘-ial’)] = [taːrʃɔʃ] (‘social’)
- [taːrʃɔʃ + ʃaːg (‘-ness’)] = [taːrʃɔʃaːg] (‘society’)
- [köz (‘place’) + taːrʃɔʃaːg] = [köztaːrʃɔʃaːg] (‘republic’)
- [nép (‘people’) + köztaːrʃɔʃaːg] = [népköztaːrʃɔʃaːg] (‘people’s republic’)
- [népköztaːrʃɔʃaːg + utsɔ (‘street’)] = [népköztaːrʃɔʃaːgutsɔ] (‘The Street of the People’s Republic’)

Latin: A Fusional Language

(Indo-European - Classical Language of the Roman Empire)

moneō	‘I am advising’
monēs	‘you(sg) are advising’
monet	‘(s)he is advising’
monēmus	‘we are advising’
monētis	‘you(pl) are advising’
monent	‘they are advising’

[-o] ‘1st, sg. pres. tense’
[-s] ‘2nd, sg. pres. tense’
[-t] ‘3rd, sg. pres. tense’
[-mus] ‘1st pl. pres. tense’
[-tis] ‘2nd pl. pres. tense’
[-nt] ‘3rd, pl. pres. tense’

Polysynthetic Languages

An example from **Chukchi** (Chukotko-Kamchatkan – 16,000 speakers)

θəmeyŋəlevtpəɣtərkən

t-ə-meyŋ-ə-levt-pəɣt-ə-rkən

1.SG.SUBJ-great-head-hurt-PRES.1

'I have a fierce headache.' (Skorik 1961: 102)

θəmeyŋəlevtpəɣtərkən has a 5:1 morpheme-to-word ratio with 3 incorporated lexical morphemes (meyŋ 'great', levt 'head', pəɣt 'ache').

Polysynthetic Languages

Two words of **Sora** (Munda (Austro-Asiatic) - 310,000):

pɔ-	poʊŋ-	koʊŋ-	t-	am
stab	belly	knife	non-past	you(sg.)

"(Someone) will stab you with a knife in (your) belly."

ɲɛn-	əʤ-	ʤa-	dar-	si-	əm
I	Not	receive	cooked-rice	hand	you(sg.)

"I will not receive cooked rice from your hands."

Note the words:
si-i "hand"; kondi "knife"

Do all languages with morphology express the same distinctions?

Morpheme Diversity

Hindi (Indo-European - 181,700,000) Causatives:

bənnaː ‘to be made’; bənaːnaː ‘to make (something)’; bənvaːnaː ‘to make (someone) make (something)’.

pəknaː ‘to be cooking’; pəkaːnaː ‘to cook (something)’; pəkvaːnaː ‘to make (someone) cook (something)’.

Saṃskṛt (IE - Classical language of ancient India) Desideratives:

pibaːti	‘he drinks’	piːpaːsati	‘he wants to drink’
jiːvati	‘he lives’	jiːjiːviʃati	‘he wants to live’

Noun classes: Swahili

(Bantu (Niger-Congo) - 800,000 native speakers; over 30,000,000 L2 users)

class	semantics	prefix	singular	gloss	plural	gloss
1,2	persons	m-/mu-, wa-	mtu	person	watu	persons
3,4	trees, natural forces	m-/mu-, mi-	mti	tree	miti	trees
5,6	groups, aug	∅/ji-, ma-	jicho	eye	macho	eyes
7,8	artifacts, dim	ki-, vi-	kisu	knife	visu	knives
9,10	animals, loanwords, other	∅/n-, ∅/n-	ndoto	dream	ndoto	dreams
11,12	extension	u-, ∅/n-	ua	fence, yard	nyua	fences
14	abstraction	u-	utoto	childhood	–	

Noun class prefixes mark singular and plural as well. Verbs contain agreement affixes:

- **wa**toto **wa**dogo **wa**meanguka
  “the small children fell.”
- **ki**tabu **ki**dogo **ki**meanguka “the small book fell.”
- **vi**tabu **vi**dogo **vi**meanguka “the small books fell.”
- **wa**toto **wa**dogo **wa**na **ki**taka **ki**tabu
  “the small children want the book.”

Allomorphs: The English Noun Plural Morpheme

CONTEXT	ALLOMORPH
baby, bag, hood, eye, hive	z
book, cat, caps, proof	s
crutch, garage, glass, buzz	əz

Phonological Rules:
The English Noun Plural Morpheme

	/bebi+z/	/bʊk+z/	/glæs+z/
Voicing Assimilation	–	[bʊk+s]	–
ə-Epenthesis	–	–	[glæs+əz]
	[bebi+z]	[bʊk+s]	[glæs+əz]

Exceptions

SINGULAR	PLURAL
man	men
woman	women
child	children
ox	oxen
tooth	teeth
foot	feet
sheep	sheep
deer	deer
fish	fish

Organizing Principle:
Exceptions (apavāda) block General Rule (utsarga)

Beyond Concatenation

- fan-ta-stic
- fan-freakin-tastic <= Infixation of “freakin” morpheme

  *fantas-freakin-tic
- Mis-sis-sip-pi
- Missi-freakin-ssippi

  *Mis-freakin-sissippi

  *Mississip-freakin-pi

- **Bound Morphemes**: cannot occur on their own as full words (-**s** in dogs; **de-** in detoxify; -**ness** in happiness; **cran-** in cranberry)
- **Free Morphemes**: can occur as separate words (**dog**; **walk**; **berry**; **yes**)

- **Zero Derivation (Conversion):** Building a different word (stem) without changing the phonology.

- ADJ → NOUN
- NOUN → VERB
- More Examples??

Ambiguity

- unusable
- prefix un-
- verb stem use
- suffix -able
- [un + [use + able]] (*unuse)

- Don’t store your money in that box, it’s unlockable.
  [un + [lock + able]]
- Now that we have the right key, the box is finally unlockable.
  [[un + lock] + able]

Morphological Vowel Mutation

- swim swam swum
- drink / drank / drunk
- begin / began / begun
- sit/sat; win/won; come/came; run/ran; shine/shone; find/found...
- wear / wore / worn (combination)

- A small number of English noun plurals also have internal changes: foot/feet; mouse/mice; man/men
- ‘Nonconcatenative’ Morphology

Arabic

FORM	MEANING	PATTERN
kataba	to write	CaCaCa
ʔaktaba	to cause to write	ʔaCCaCa
kaatib	writing	CaaCiC
kitaab	a book	CiCaaC
kutub	boo	CuCuC
kitaabah	writing profession	CiCaaCah
kattaab	author	CaCCaaC
miktaab	writing instrument	miCCaaC

Arabic

FORM	MEANING	PATTERN
kataba	he wrote	CaCaCa
katabna	we wrote	CaCaCna
katabuu	they wrote	CaCaCuu
yaktubu	he writes	yaCCuCu
naktubu	we write	naCCuCu
yaktabuuna	they write	yaCCaCuuna
sayaktubu	he will write	sayaCCuCu
sanaktubu	we will write	sanaCCuCu
sayaktabuuna	they will write	sayaCCaCuuna

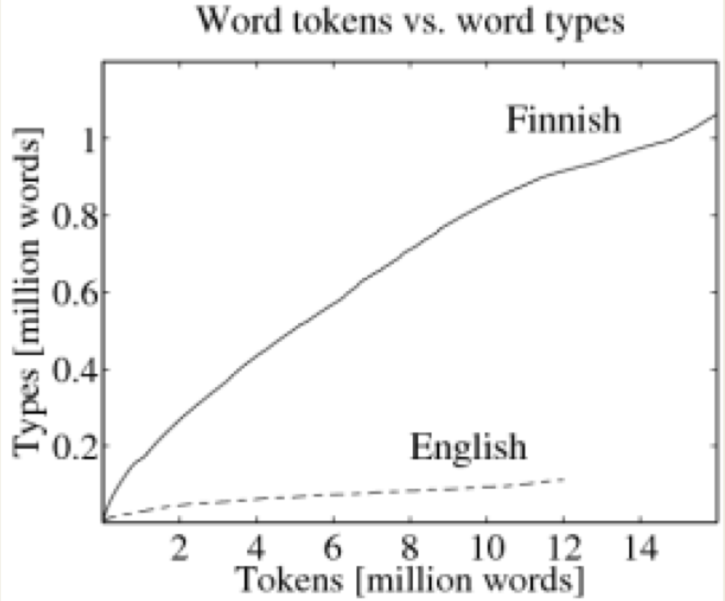
Morphology
for Machine Translation

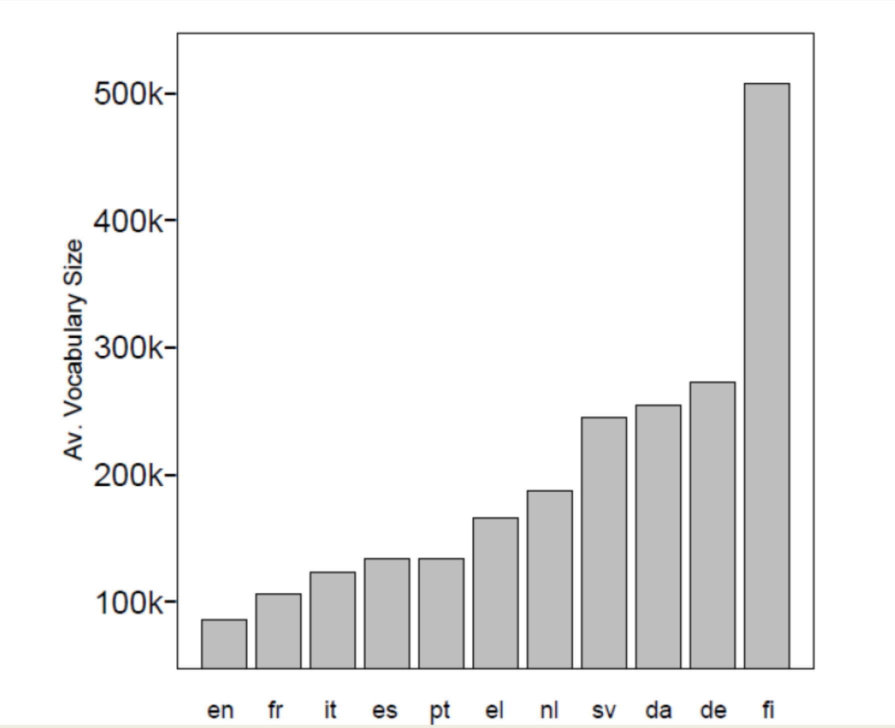
Long distance agreement error

REF: Maria is buying her first house

MT: Maria is buying his first house

Sparsity





tietä+isi+mme

know+would+we

Creutz et al. 2005

high → low inflected

- Preprocessing techniques
  - Segmentation approaches

“easy” task

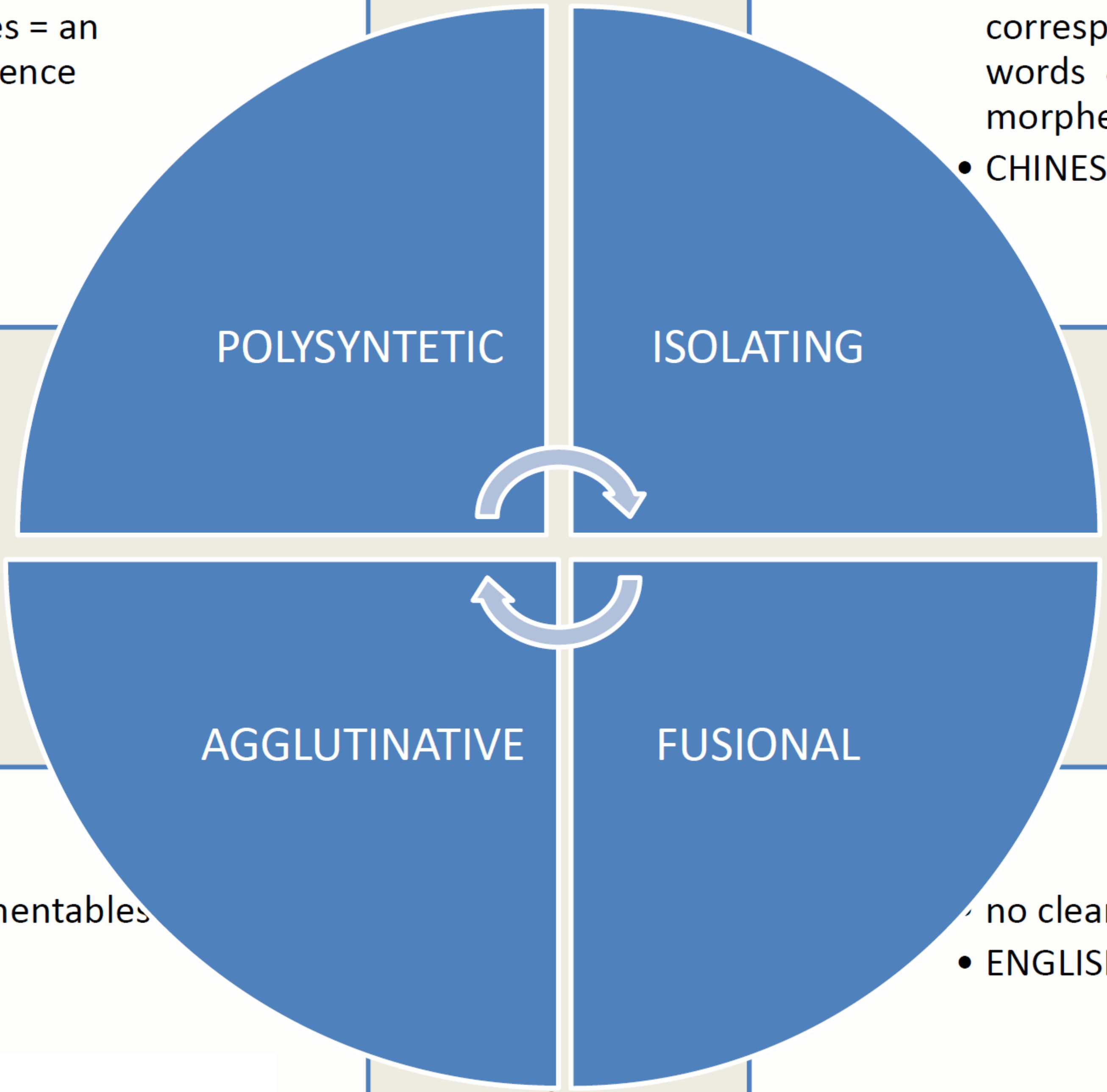
from big to small space

low → high inflected

- Postprocessing techniques
  - Generation
  - Enriching models

difficult task

from small to big space



- a word of several morphemes = an entire sentence
- INUIT

POLYSYNTETIC

ISOLATING

- one-to-one correspondance words & morphemes.
- CHINESE

AGGLUTINATIVE

FUSIONAL

- easily segmentables
- TURKISH

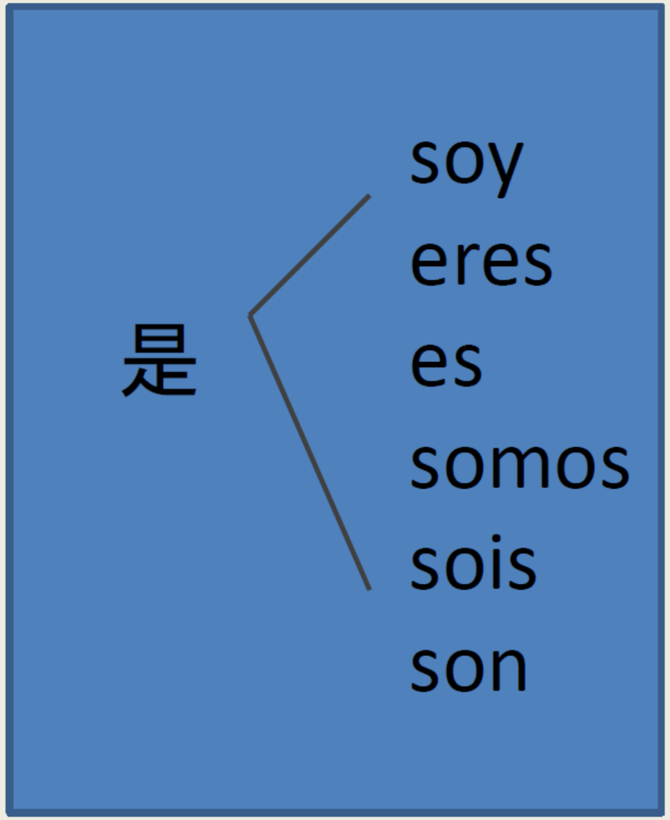
- no clear boundaries
- ENGLISH

isolating ↔ fusional/agglutinative

- Isolating language

  是

- High-inflected language
  - Yo soy -Nosotros somos
  - Tu eres -Vosotros sois
  - Él es -Ellos son



Chinese ↔ Spanish

Language-dependent segmentation

- English into Spanish/Catalan task:
  - Treatment of verbs: identify (by means of POS) pronoun+verb sequence and splice these two words into one,
    - » you go --- PRP VBP --- you_go
- Spanish/Catalan into English task:
  - split contractions (e.g. del = de + el, al =a +el)

Ueffing et al. 2003

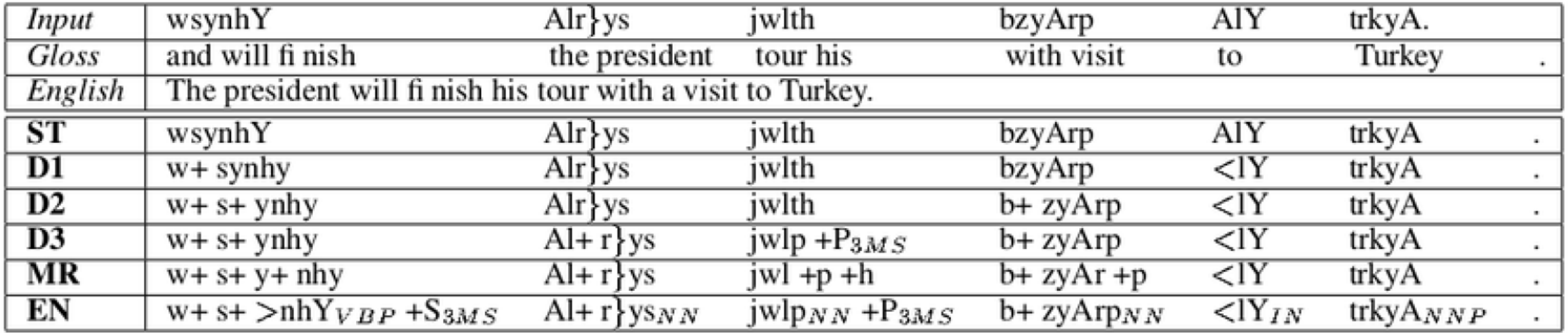
Language-dependent segmentation

- Arabic-to-English task.

TOK	
ST	Splitting off punctuation and numbers
D1	Declitization (w+, f+)
D2	Declitization (D1+ l+, k+, b+, s+)
D3	Declitization (D1,D2, Al+)
MR	Stem + affixival morphemes
EN	English-like

Language-dependent segmentation

- Arabic-to-English task.



Input	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA.	
Gloss	and will finish	the president	tour his	with visit	to	Turkey	.
English	The president will finish his tour with a visit to Turkey.						
ST	wsynhY	Alr}ys	jwlth	bzyArp	AlY	trkyA	.
D1	w+ synhy	Alr}ys	jwlth	bzyArp	<lY	trkyA	.
D2	w+ s+ ynhy	Alr}ys	jwlth	b+ zyArp	<lY	trkyA	.
D3	w+ s+ ynhy	Al+ r}ys	jwlp +P$_{3MS}$	b+ zyArp	<lY	trkyA	.
MR	w+ s+ y+ nhy	Al+ r}ys	jwl +p +h	b+ zyAr +p	<lY	trkyA	.
EN	w+ s+ >nhY$_{VBP}$ +S$_{3MS}$	Al+ r}ys$_{NN}$	jwlp$_{NN}$ +P$_{3MS}$	b+ zyArp$_{NN}$	<lY$_{IN}$	trkyA$_{NNP}$	.

  - Small data set: English-like tokenization
  - Large data set: splitting only some clíticS

Language-independent segmentation

- Morfessor is a method for finding morpheme-like units of a language in an unsupervised manner.
  - Minimum Description Length

Example of segmentation:

affectionate	affect+ion+ate

Common morphological operations

- AFFIXATION: *nation + al*
- COMPOUNDING: *sun+glasses*
- REDUPLICATION: *bye-bye*
- INTERNAL CHANGE: *rang [instead of ringed]*
- SUPPLETION: *went [past of go*]
- BLENDING*: motel [motor+hotel]*

Factored translation models

- Factored translation models are an extension to phrase-based models where every word is substituted by a vector of factors.

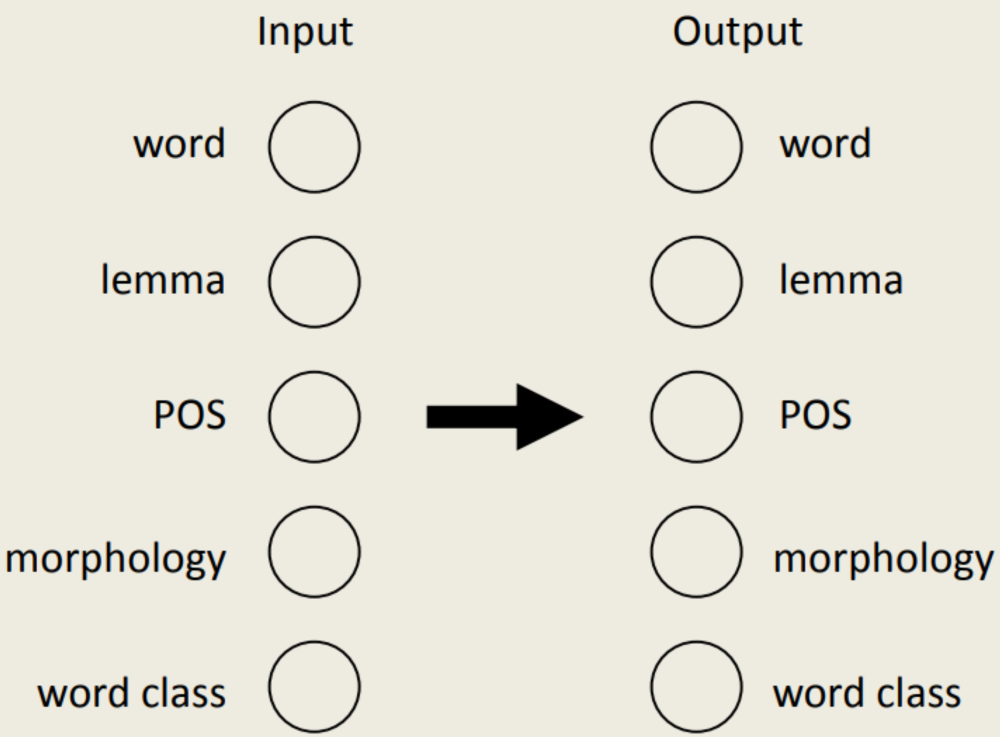
(word) =⇒ (word, lemma, PoS, morphology, ...)

- The translation is now a combination of pure translation (T) and generation (G) steps:

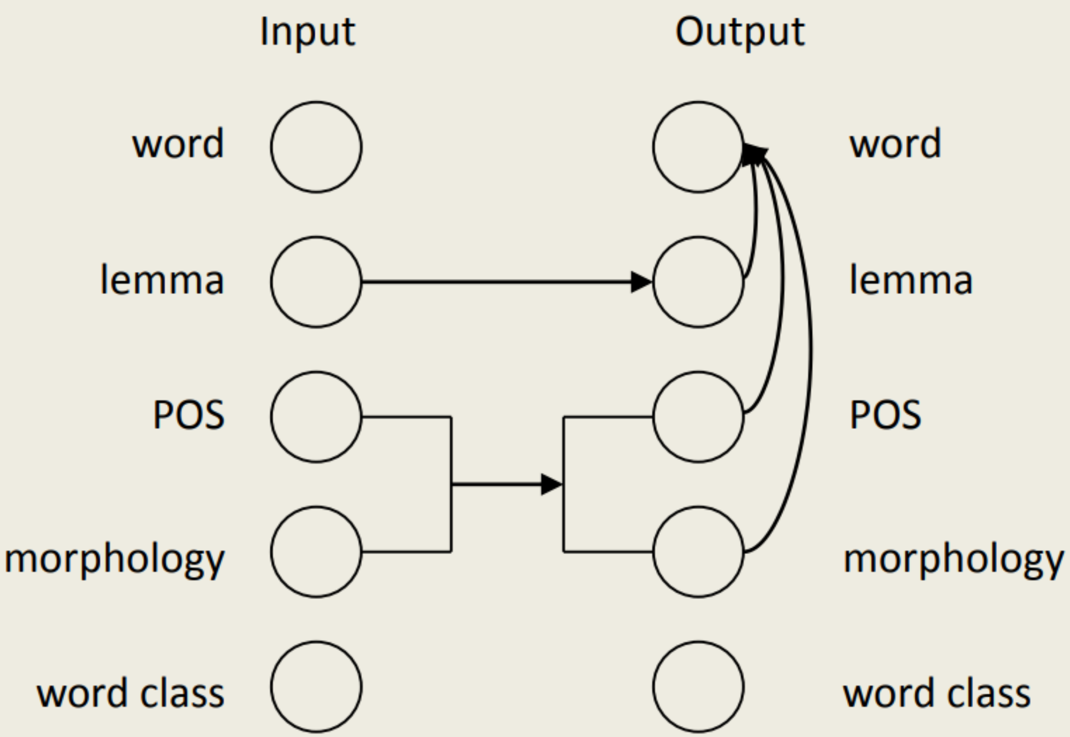
$$\begin{tikzcd}
\texttt{lemma}_f \arrow[d, "\mathsf{T}"] & \texttt{PoS}_f \arrow[d, "\mathsf{T}"] & \texttt{morphology}_f \arrow[d, "\mathsf{T}"] & \texttt{word}_f \\
\texttt{lemma}_e & \texttt{PoS}_e & \texttt{morphology}_e \arrow[r, "\mathsf{G}"] & \texttt{word}_e
\end{tikzcd}$$

Factored translation models

Factored Representation



Factored Model: transfer and generation



Factored translation models

What differs in factored translation models
(as compared to standard phrase-based models)

- The parallel corpus must be annotated beforehand.
- Extra language models for every factor can also be used.
- Translation steps are accomplished in a similar way.
- Generation steps imply a training only on the target side of the corpus.
- Models corresponding to the different factors and components are combined in a log-linear fashion.

PoS verb morphology simplification

Type	Text
Plain target	La Comisión puede llegar a paralizar el programa
Lemma + PoS	La Comisión VMIP3S0[poder] llegar a paralizar el programa
Lemma+PoS Generalized	La Comisión VMIpn0[poder] llegar a paralizar el programa

Learning Unseen Forms

Small Parallel Data

Source	Target	Target Lemma
A cat chased	**kočka honila...**	**kočka honit...**
I saw a cat	**kočku vidět**	**být kočka**
I read about a dog	četl jsem o psovi	číst být o pes

Large Monolingual Data:

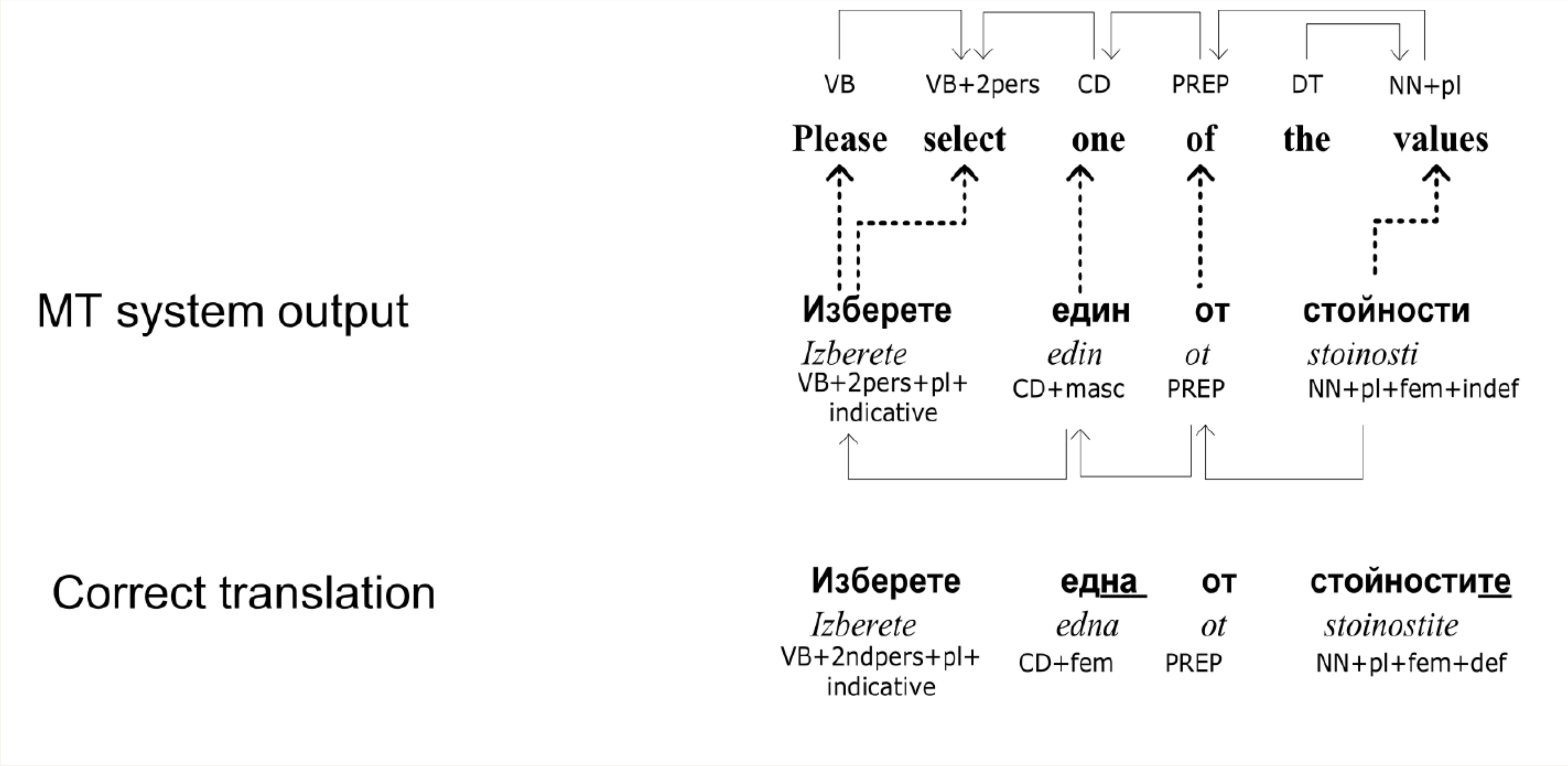
Source	Target	Target Lemma
?	**četl jsem o kočce**	**číst být o kočka**

I read about a cat ¬ Use reverse translation backed-off by lemmas

- Learned a new phrase (**o kočce**) including a form never seen in parallel data (**kočce**).

Discriminative selection models

- Better lexical selection, especially for morphologically complex languages



VB VB+2pers CD PREP DT NN+pl

Please select one of the values

MT system output

Изберете *Izberete* VB+2pers+pl+ indicative

един *edin* CD+masc

от *ot* PREP

стойности *stoinosti* NN+pl+fem+indef

Correct translation

Изберете *Izberete* VB+2ndpers+pl+ indicative

ед<u>на</u> *edna* CD+fem

от *ot* PREP

стойности<u>те</u> *stoinostite* NN+pl+fem+def

Jeong, Toutanova, Suzuki, and Quirk 2010

Morphology at JHU

Collaborators:

Faculty: David Yarowsky, Philipp Koehn, Matt Post, Kevin Duh, Jason Eisner

Senior Researchers/Postdocs:

Christo Kirov, Garrett Nicolai, Oliver Adams, John Sylak-Glassman

PhD Students:

Winston Wu, Arya McCarthy, Ryan Cotterell,
Aaron Mueller, Huda Khayrallah, Patrick Xia

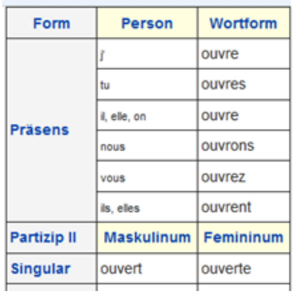
Masters/Undergraduates:

Nidhi Vyas, John Hewitt, Roger Que, James Scharf

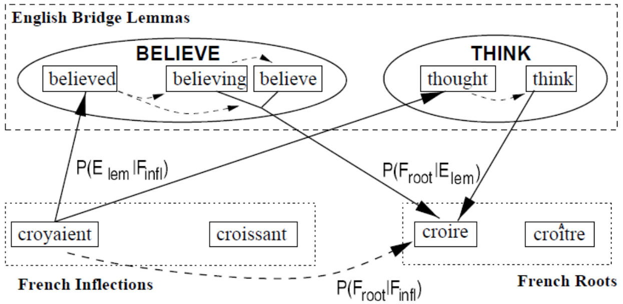
Dylan Lewis, Lawrence Wolf-Sarkin, ++

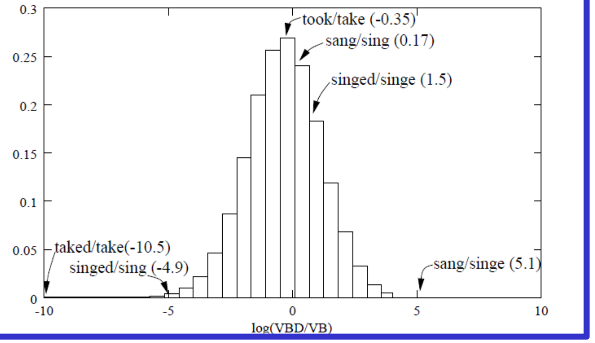
Multi-Source/Multi-Stage Morphology Learning:

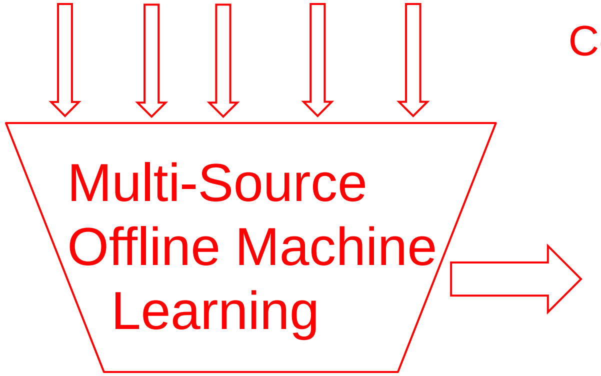
- Currently available supervised data (e.g. Wiktionary)
- Elicited paradigms (professional translators, Mturk)
- Seed data from grammars, ITG, linguistic universals
- Bilingual projection (e.g. from aligned Bibles)
- Monolingual contextual/distributional statistics



Form	Person	Wortform
Präsens	j'	ouvre
	tu	ouvres
	il, elle, on	ouvre
	nous	ouvrons
	vous	ouvrez
	ils, elles	ouvrent
Partizip II	Maskulinum	Femininum
Singular	ouvert	ouverte

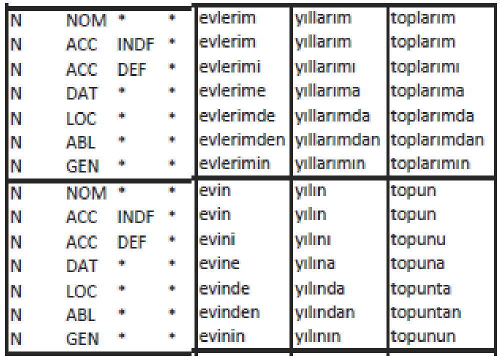






Multi-Source Offline Machine Learning

Complete Learned Paradigms

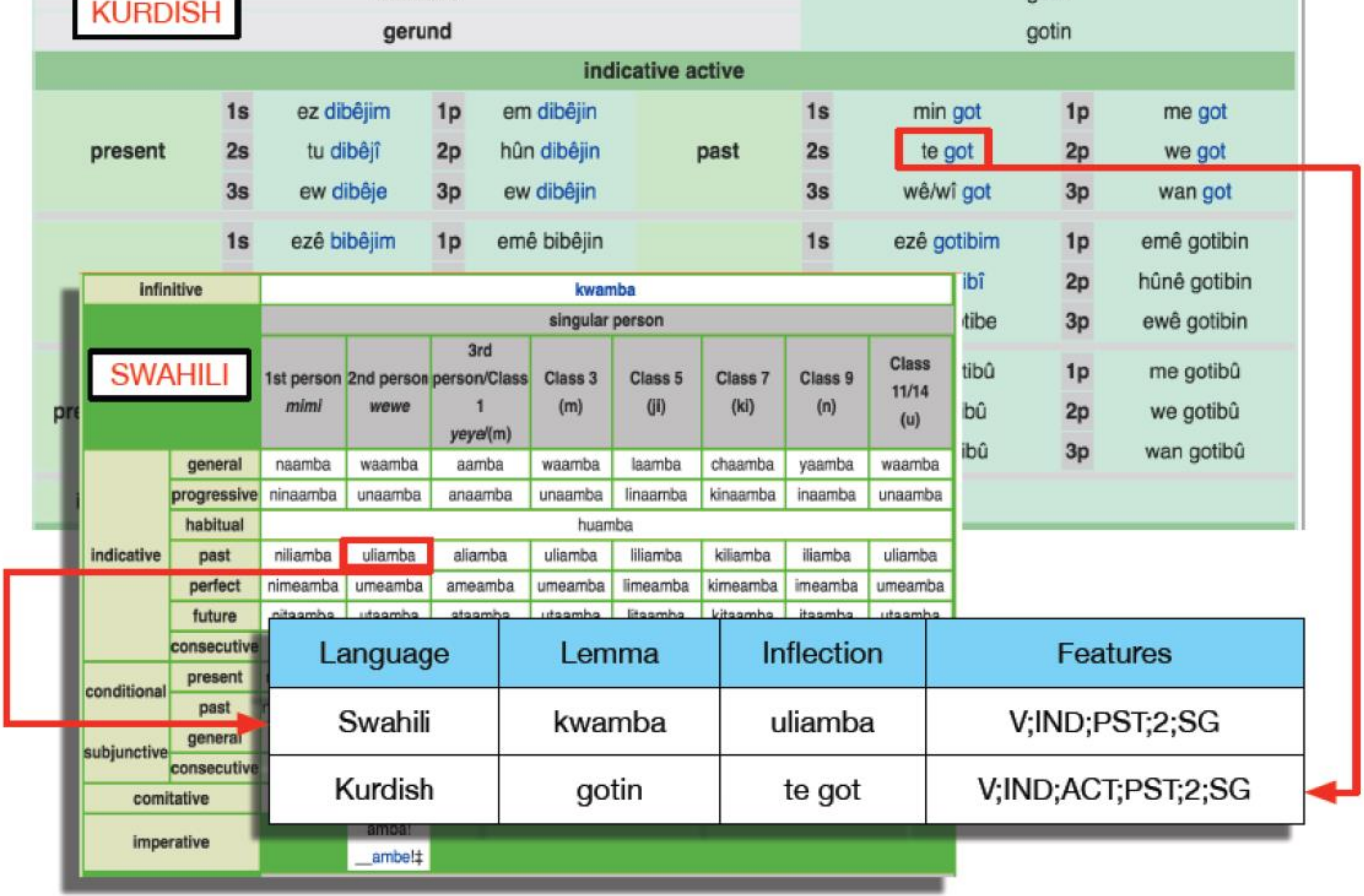


N	NOM	*	*	evlerim	yıllarım	toplarım
N	ACC	INDF	*	evlerim	yıllarım	toplarım
N	ACC	DEF	*	evlerimi	yıllarımı	toplarımı
N	DAT	*	*	evlerime	yıllarıma	toplarıma
N	LOC	*	*	evlerimde	yıllarımda	toplarımda
N	ABL	*	*	evlerimden	yıllarımdan	toplarımdan
N	GEN	*	*	evlerimin	yıllarımın	toplarımın
N	NOM	*	*	evin	yılın	topun
N	ACC	INDF	*	evin	yılın	topun
N	ACC	DEF	*	evini	yılını	topunu
N	DAT	*	*	evine	yılına	topuna
N	LOC	*	*	evinde	yılında	topunta
N	ABL	*	*	evinden	yılından	topuntan
N	GEN	*	*	evinin	yılının	topunun

Human Vetting/Improvement

Run-time Executables and importable hash tables

>> DO THIS FOR 300-1600 WORLD LANGUAGES!



KURDISH

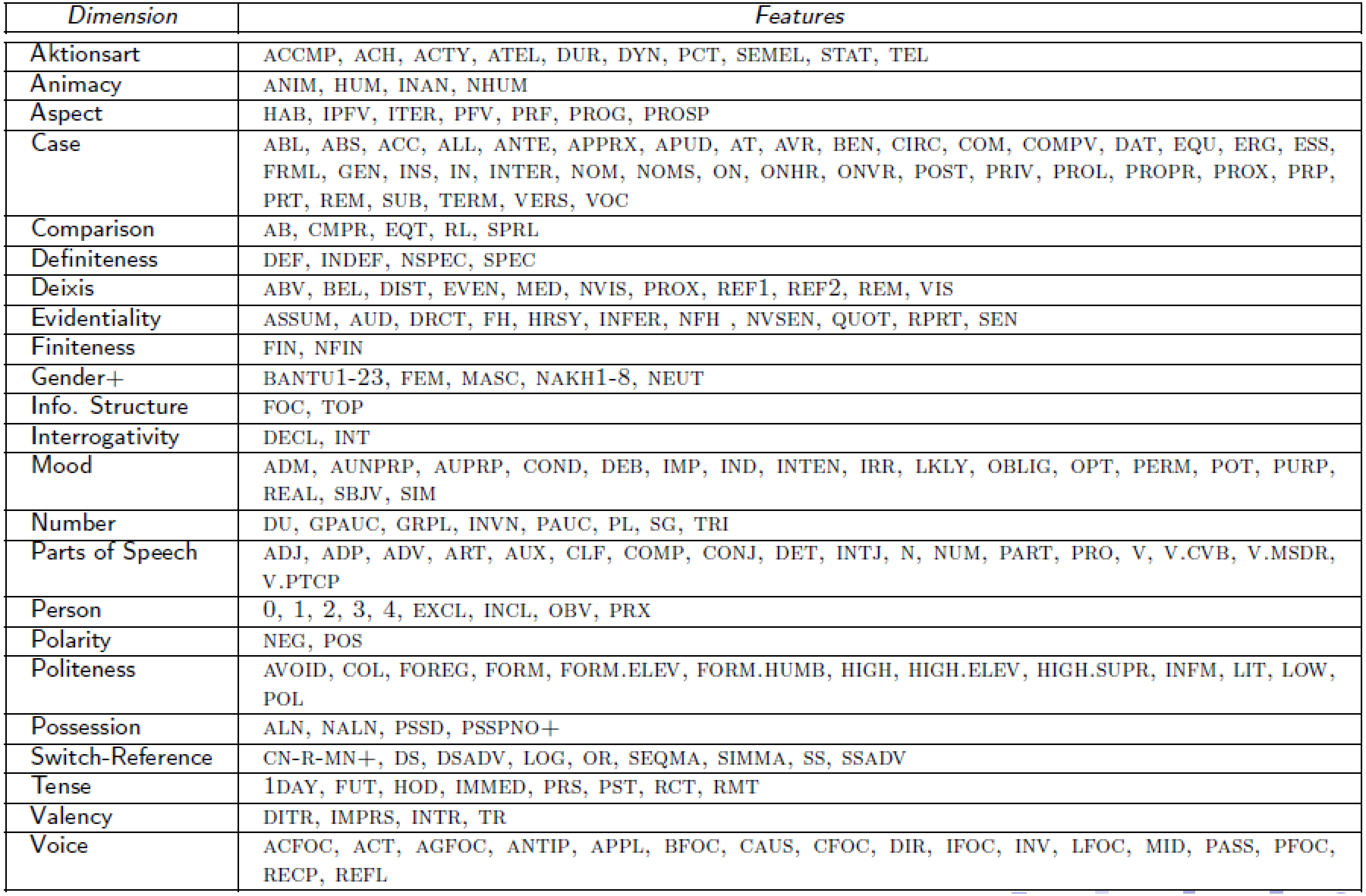
gerund					gotin				
indicative active									
present	1s	ez dibêjim	1p	em dibêjin	past	1s	min got	1p	me got
	2s	tu dibêjî	2p	hûn dibêjin		2s	te got	2p	we got
	3s	ew dibêje	3p	ew dibêjin		3s	wê/wî got	3p	wan got
	1s	ezê bibêjim	1p	emê bibêjin		1s	ezê gotibim	1p	emê gotibin
							ibî	2p	hûnê gotibin
							tibe	3p	ewê gotibin
pr							tibû	1p	me gotibû
							bû	2p	we gotibû
							ibû	3p	wan gotibû

SWAHILI

infinitive		kwamba							
		singular person							
		1st person *mimi*	2nd person *wewe*	3rd person/Class 1 *yeye*/(m)	Class 3 (m)	Class 5 (ji)	Class 7 (ki)	Class 9 (n)	Class 11/14 (u)
indicative	general	naamba	waamba	aamba	waamba	laamba	chaamba	yaamba	waamba
	progressive	ninaamba	unaamba	anaamba	unaamba	linaamba	kinaamba	inaamba	unaamba
	habitual	huamba							
	past	niliamba	uliamba	aliamba	uliamba	liliamba	kiliamba	iliamba	uliamba
	perfect	nimeamba	umeamba	ameamba	umeamba	limeamba	kimeamba	imeamba	umeamba
	future	nitaamba	utaamba	ataamba	utaamba	litaamba	kitaamba	itaamba	utaamba
	consecutive								
conditional	present								
	past								
subjunctive	general								
	consecutive								
comitative									
imperative			amba! __ambe!‡						

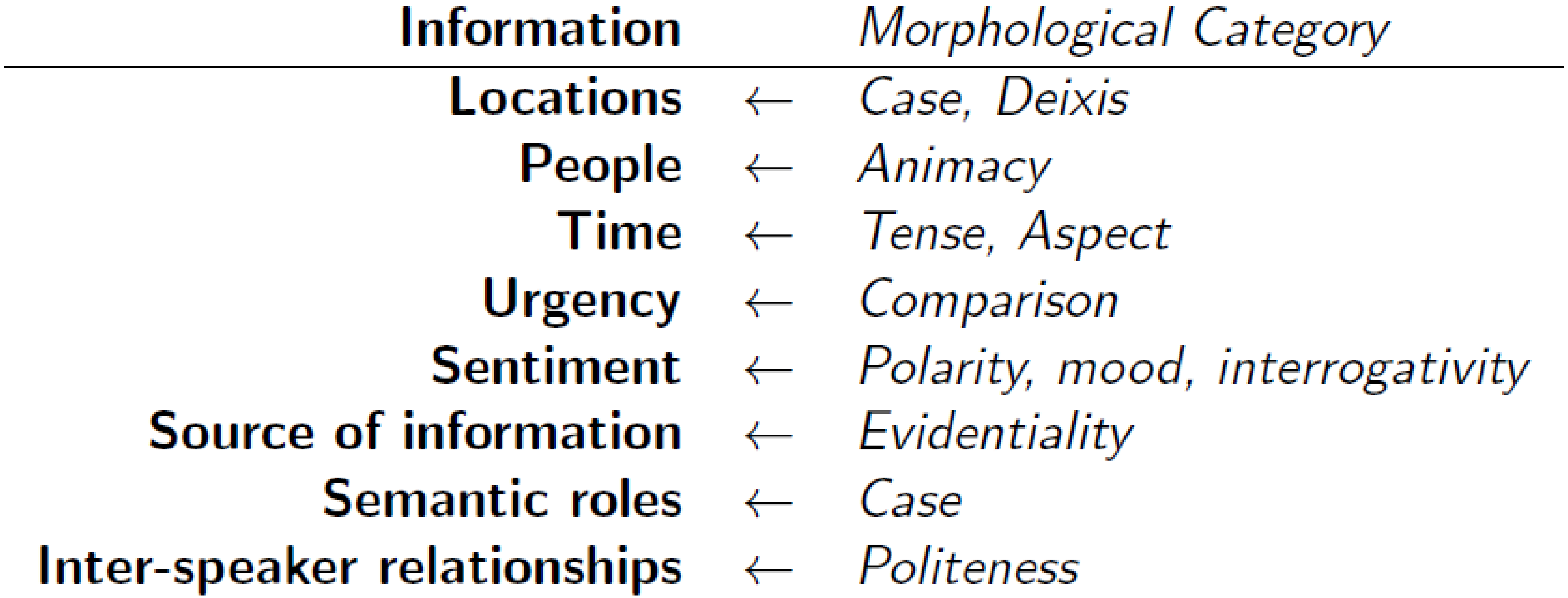
Language	Lemma	Inflection	Features
Swahili	kwamba	uliamba	V;IND;PST;2;SG
Kurdish	gotin	te got	V;IND;ACT;PST;2;SG

UniMorph Feature Schema (dimensions of meaning)



Dimension	*Features*
Aktionsart	ACCMP, ACH, ACTY, ATEL, DUR, DYN, PCT, SEMEL, STAT, TEL
Animacy	ANIM, HUM, INAN, NHUM
Aspect	HAB, IPFV, ITER, PFV, PRF, PROG, PROSP
Case	ABL, ABS, ACC, ALL, ANTE, APPRX, APUD, AT, AVR, BEN, CIRC, COM, COMPV, DAT, EQU, ERG, ESS, FRML, GEN, INS, IN, INTER, NOM, NOMS, ON, ONHR, ONVR, POST, PRIV, PROL, PROPR, PROX, PRP, PRT, REM, SUB, TERM, VERS, VOC
Comparison	AB, CMPR, EQT, RL, SPRL
Definiteness	DEF, INDEF, NSPEC, SPEC
Deixis	ABV, BEL, DIST, EVEN, MED, NVIS, PROX, REF1, REF2, REM, VIS
Evidentiality	ASSUM, AUD, DRCT, FH, HRSY, INFER, NFH , NVSEN, QUOT, RPRT, SEN
Finiteness	FIN, NFIN
Gender+	BANTU1-23, FEM, MASC, NAKH1-8, NEUT
Info. Structure	FOC, TOP
Interrogativity	DECL, INT
Mood	ADM, AUNPRP, AUPRP, COND, DEB, IMP, IND, INTEN, IRR, LKLY, OBLIG, OPT, PERM, POT, PURP, REAL, SBJV, SIM
Number	DU, GPAUC, GRPL, INVN, PAUC, PL, SG, TRI
Parts of Speech	ADJ, ADP, ADV, ART, AUX, CLF, COMP, CONJ, DET, INTJ, N, NUM, PART, PRO, V, V.CVB, V.MSDR, V.PTCP
Person	0, 1, 2, 3, 4, EXCL, INCL, OBV, PRX
Polarity	NEG, POS
Politeness	AVOID, COL, FOREG, FORM, FORM.ELEV, FORM.HUMB, HIGH, HIGH.ELEV, HIGH.SUPR, INFM, LIT, LOW, POL
Possession	ALN, NALN, PSSD, PSSPNO+
Switch-Reference	CN-R-MN+, DS, DSADV, LOG, OR, SEQMA, SIMMA, SS, SSADV
Tense	1DAY, FUT, HOD, IMMED, PRS, PST, RCT, RMT
Valency	DITR, IMPRS, INTR, TR
Voice	ACFOC, ACT, AGFOC, ANTIP, APPL, BFOC, CAUS, CFOC, DIR, IFOC, INV, LFOC, MID, PASS, PFOC, RECP, REFL

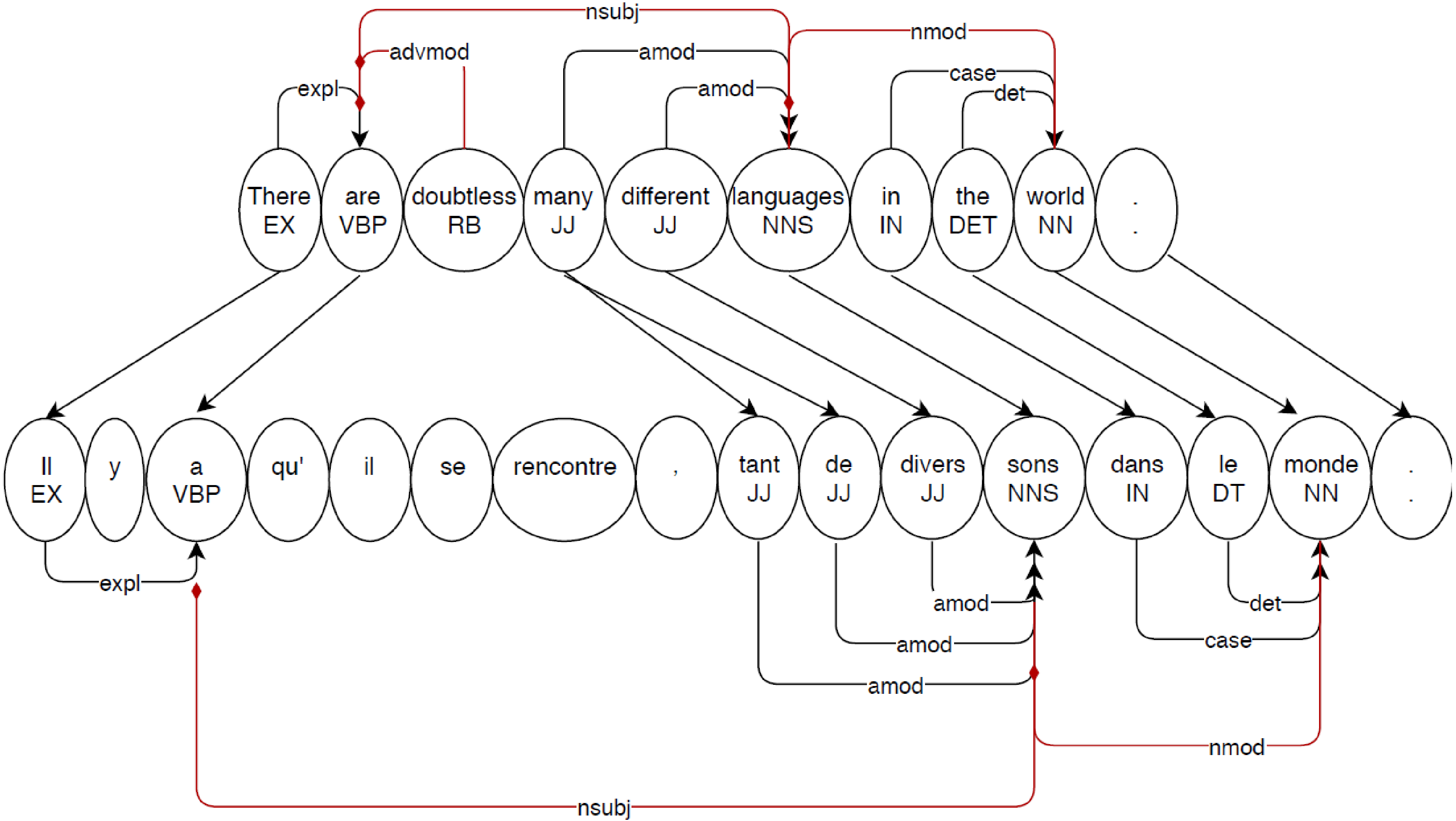
Example UniMorph uses in Information Extraction:



Information		*Morphological Category*
Locations	←	*Case, Deixis*
People	←	*Animacy*
Time	←	*Tense, Aspect*
Urgency	←	*Comparison*
Sentiment	←	*Polarity, mood, interrogativity*
Source of information	←	*Evidentiality*
Semantic roles	←	*Case*
Inter-speaker relationships	←	*Politeness*

Projection of POS tags and Dependency Parses

(English semantic roles identify target cases; nsubj dependencies give Person/Number)



Example Unimorph Output:

Tables of English phrasal translations of inflected forms

INPUT

OUTPUT

SpInf	SpRoot	Unimorph Vector	English Template	English phrasal inflection
comía	comer	V;IPFV;PST;1;SG	I was VBG	I was eating
comías	comer	V;IPFV;PST;2;SG;INFM	you were VBG	you were eating
comías	comer	V;IPFV;PST;2;SG;FORM	you were VBG	you were eating
comía	comer	V;IPFV;PST;3;SG	he/she/it was VBG	he/she/it was eating
comíamos	comer	V;IPFV;PST;1;PL	we were VBG	we were eating
comíais	comer	V;IPFV;PST;2;PL;INFM	you all were VBG	you all were eating
comíais	comer	V;IPFV;PST;2;PL	you all were VBG	you all were eating
comían	comer	V;IPFV;PST;3;PL	they were VBG	they were eating
hablaba	hablar	V;IPFV;PST;1;SG	I was VBG	I was speaking
hablabas	hablar	V;IPFV;PST;2;SG;INFM	you were VBG	you were speaking
hablabas	hablar	V;IPFV;PST;2;SG;FORM	you were VBG	you were speaking
hablaba	hablar	V;IPFV;PST;3;SG	he/she/it was VBG	he/she/it was speaking
hablábamos	hablar	V;IPFV;PST;1;PL	we were VBG	we were speaking
hablais	hablar	V;IPFV;PST;2;PL;INFM	you all were VBG	you all were speaking
hablais	hablar	V;IPFV;PST;2;PL	you all were VBG	you all were speaking
hablaban	hablar	V;IPFV;PST;3;PL	they were VBG	they were speaking

GitHub distribution of Trained Morphological Analyzers AND generators for 903+ languages!

(will soon be 1100+)

Diverse detailed inflectional morphology

Nouns: sg/pl and case(nom/acc/dat/gen/loc/other)

Verbs: tense(pst/prs/fut) +person/number(1SG,1PL,2..)

Adjectives: person/number/case/gender in progress

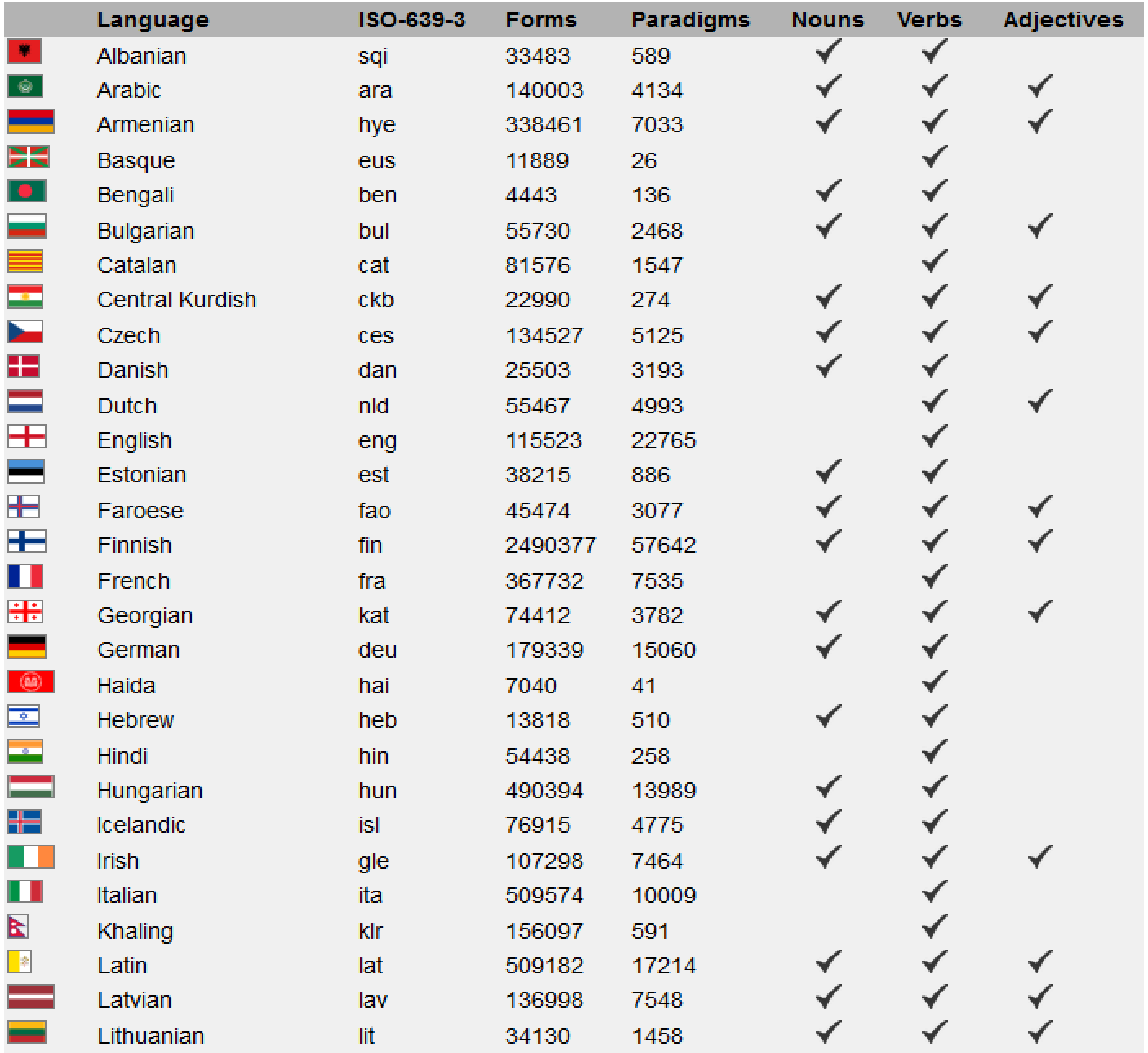
Analysis mode:

```
python analyze.py   -i Inflected-Zapotec.txt  -a Zapotec.analyses  -l zap  -d Zapotec-lemma-list
```

Generation mode:

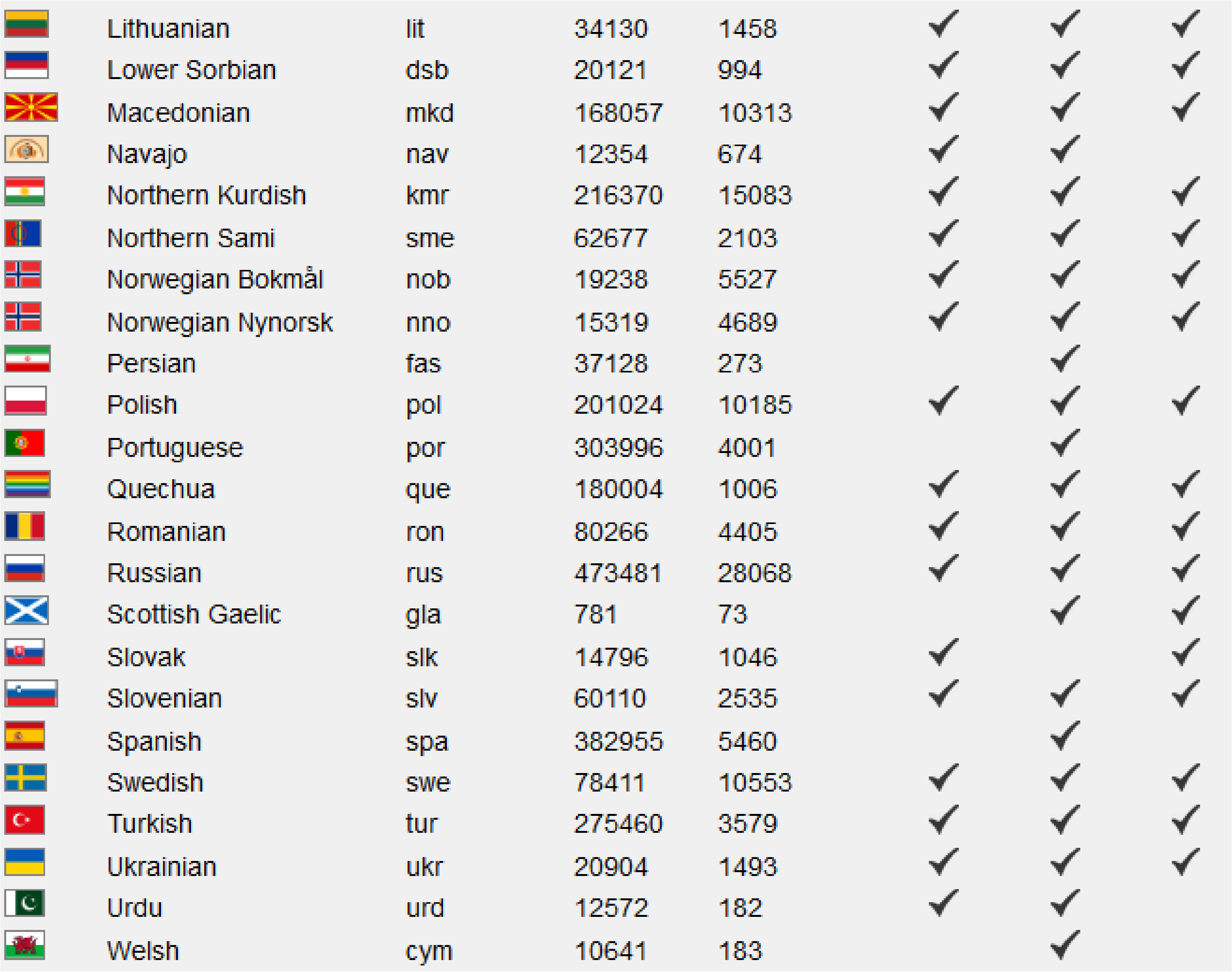
```
python analyze.py   -i Zapotec-lemma-list  -g -a Zapotec.generation  -l zap  -d Zapotec-corpus-words
```

UniMorph (example of currently released languages)



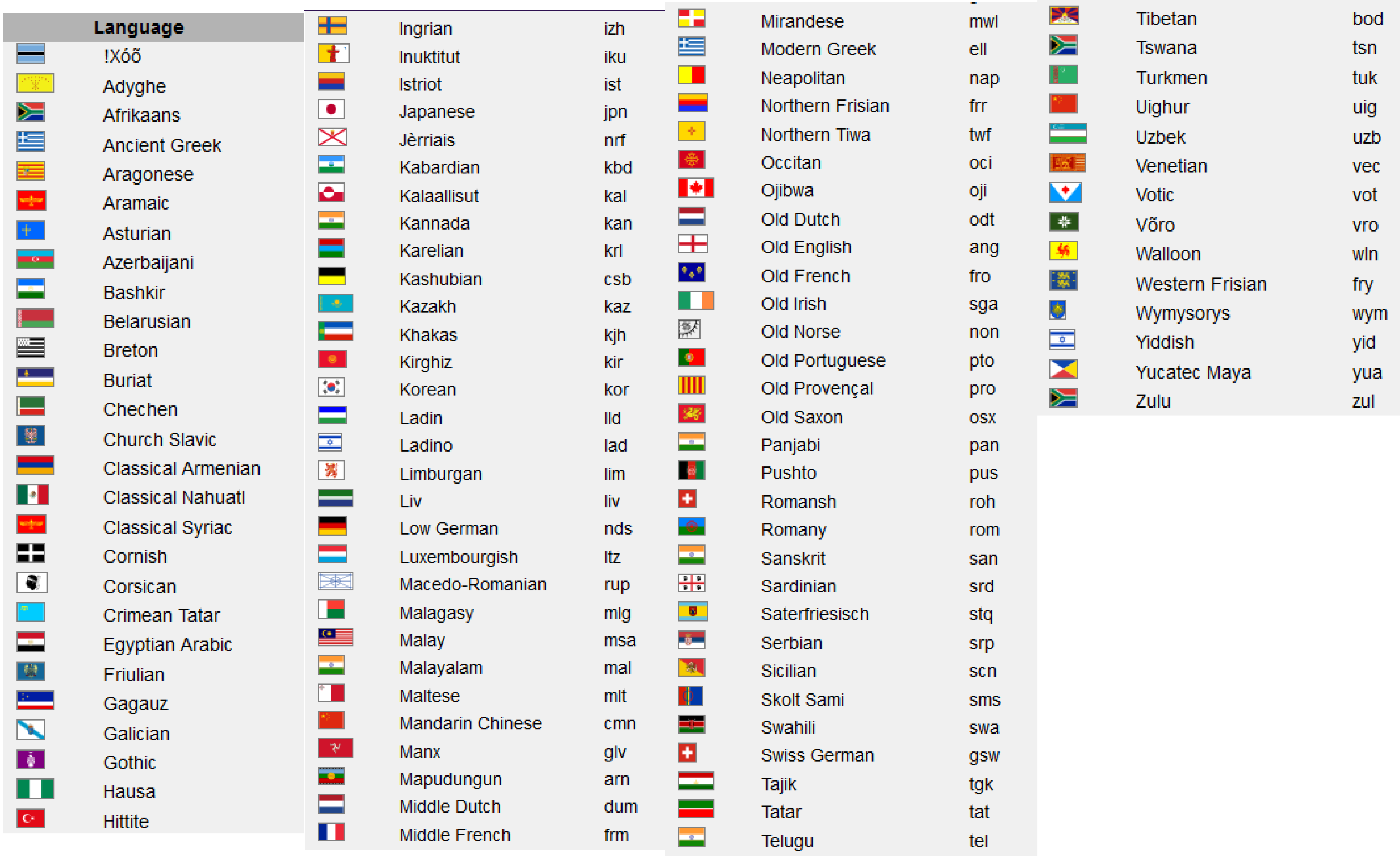
Language	ISO-639-3	Forms	Paradigms	Nouns	Verbs	Adjectives
Albanian	sqi	33483	589	✓	✓	
Arabic	ara	140003	4134	✓	✓	✓
Armenian	hye	338461	7033	✓	✓	✓
Basque	eus	11889	26		✓	
Bengali	ben	4443	136	✓	✓	
Bulgarian	bul	55730	2468	✓	✓	✓
Catalan	cat	81576	1547		✓	
Central Kurdish	ckb	22990	274	✓	✓	✓
Czech	ces	134527	5125	✓	✓	✓
Danish	dan	25503	3193	✓	✓	
Dutch	nld	55467	4993		✓	✓
English	eng	115523	22765		✓	
Estonian	est	38215	886	✓	✓	
Faroese	fao	45474	3077	✓	✓	✓
Finnish	fin	2490377	57642	✓	✓	✓
French	fra	367732	7535		✓	
Georgian	kat	74412	3782	✓	✓	✓
German	deu	179339	15060	✓	✓	
Haida	hai	7040	41		✓	
Hebrew	heb	13818	510	✓	✓	
Hindi	hin	54438	258		✓	
Hungarian	hun	490394	13989	✓	✓	
Icelandic	isl	76915	4775	✓	✓	
Irish	gle	107298	7464	✓	✓	✓
Italian	ita	509574	10009		✓	
Khaling	klr	156097	591		✓	
Latin	lat	509182	17214	✓	✓	✓
Latvian	lav	136998	7548	✓	✓	✓
Lithuanian	lit	34130	1458	✓	✓	✓

UniMorph Languages (continued)



	Lithuanian	lit	34130	1458	✓	✓	✓
	Lower Sorbian	dsb	20121	994	✓	✓	✓
	Macedonian	mkd	168057	10313	✓	✓	✓
	Navajo	nav	12354	674	✓	✓	
	Northern Kurdish	kmr	216370	15083	✓	✓	✓
	Northern Sami	sme	62677	2103	✓	✓	✓
	Norwegian Bokmål	nob	19238	5527	✓	✓	✓
	Norwegian Nynorsk	nno	15319	4689	✓	✓	✓
	Persian	fas	37128	273		✓	
	Polish	pol	201024	10185	✓	✓	✓
	Portuguese	por	303996	4001		✓	
	Quechua	que	180004	1006	✓	✓	✓
	Romanian	ron	80266	4405	✓	✓	✓
	Russian	rus	473481	28068	✓	✓	✓
	Scottish Gaelic	gla	781	73		✓	✓
	Slovak	slk	14796	1046	✓		✓
	Slovenian	slv	60110	2535	✓	✓	✓
	Spanish	spa	382955	5460		✓	
	Swedish	swe	78411	10553	✓	✓	✓
	Turkish	tur	275460	3579	✓	✓	✓
	Ukrainian	ukr	20904	1493	✓	✓	✓
	Urdu	urd	12572	182	✓	✓	
	Welsh	cym	10641	183		✓	

UniMorph Languages (continued – page #3)



Language	
!Xóõ	
Adyghe	
Afrikaans	
Ancient Greek	
Aragonese	
Aramaic	
Asturian	
Azerbaijani	
Bashkir	
Belarusian	
Breton	
Buriat	
Chechen	
Church Slavic	
Classical Armenian	
Classical Nahuatl	
Classical Syriac	
Cornish	
Corsican	
Crimean Tatar	
Egyptian Arabic	
Friulian	
Gagauz	
Galician	
Gothic	
Hausa	
Hittite	
Ingrian	izh
Inuktitut	iku
Istriot	ist
Japanese	jpn
Jèrriais	nrf
Kabardian	kbd
Kalaallisut	kal
Kannada	kan
Karelian	krl
Kashubian	csb
Kazakh	kaz
Khakas	kjh
Kirghiz	kir
Korean	kor
Ladin	lld
Ladino	lad
Limburgan	lim
Liv	liv
Low German	nds
Luxembourgish	ltz
Macedo-Romanian	rup
Malagasy	mlg
Malay	msa
Malayalam	mal
Maltese	mlt
Mandarin Chinese	cmn
Manx	glv
Mapudungun	arn
Middle Dutch	dum
Middle French	frm
Mirandese	mwl
Modern Greek	ell
Neapolitan	nap
Northern Frisian	frr
Northern Tiwa	twf
Occitan	oci
Ojibwa	oji
Old Dutch	odt
Old English	ang
Old French	fro
Old Irish	sga
Old Norse	non
Old Portuguese	pto
Old Provençal	pro
Old Saxon	osx
Panjabi	pan
Pushto	pus
Romansh	roh
Romany	rom
Sanskrit	san
Sardinian	srd
Saterfriesisch	stq
Serbian	srp
Sicilian	scn
Skolt Sami	sms
Swahili	swa
Swiss German	gsw
Tajik	tgk
Tatar	tat
Telugu	tel
Tibetan	bod
Tswana	tsn
Turkmen	tuk
Uighur	uig
Uzbek	uzb
Venetian	vec
Votic	vot
Võro	vro
Walloon	wln
Western Frisian	fry
Wymysorys	wym
Yiddish	yid
Yucatec Maya	yua
Zulu	zul

Example Unimorph Output:

Tables of English phrasal translations of inflected forms

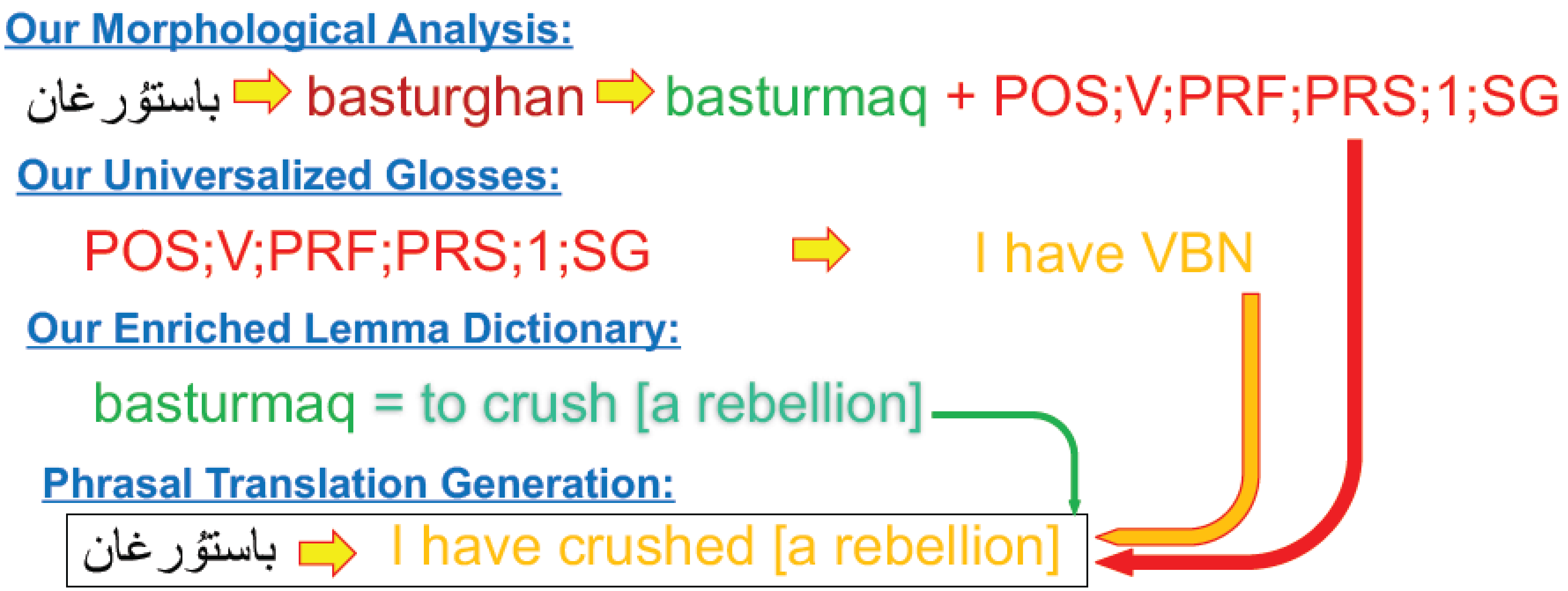
INPUT

OUTPUT

SpInf	SpRoot	Unimorph Vector	English Template	English phrasal inflection
comía	comer	V;IPFV;PST;1;SG	I was VBG	I was eating
comías	comer	V;IPFV;PST;2;SG;INFM	you were VBG	you were eating
comías	comer	V;IPFV;PST;2;SG;FORM	you were VBG	you were eating
comía	comer	V;IPFV;PST;3;SG	he/she/it was VBG	he/she/it was eating
comíamos	comer	V;IPFV;PST;1;PL	we were VBG	we were eating
comíais	comer	V;IPFV;PST;2;PL;INFM	you all were VBG	you all were eating
comíais	comer	V;IPFV;PST;2;PL	you all were VBG	you all were eating
comían	comer	V;IPFV;PST;3;PL	they were VBG	they were eating
hablaba	hablar	V;IPFV;PST;1;SG	I was VBG	I was speaking
hablabas	hablar	V;IPFV;PST;2;SG;INFM	you were VBG	you were speaking
hablabas	hablar	V;IPFV;PST;2;SG;FORM	you were VBG	you were speaking
hablaba	hablar	V;IPFV;PST;3;SG	he/she/it was VBG	he/she/it was speaking
hablábamos	hablar	V;IPFV;PST;1;PL	we were VBG	we were speaking
hablais	hablar	V;IPFV;PST;2;PL;INFM	you all were VBG	you all were speaking
hablais	hablar	V;IPFV;PST;2;PL	you all were VBG	you all were speaking
hablaban	hablar	V;IPFV;PST;3;PL	they were VBG	they were speaking

UniMorph Gloss Use for Machine Translation

- Combined universalized glosses, morphological analyses and our consensus translation lexicons to generate phrasal translations.



Our Morphological Analysis:

باستۇرغان ⇨ basturghan ⇨ basturmaq + POS;V;PRF;PRS;1;SG

Our Universalized Glosses:

POS;V;PRF;PRS;1;SG ⇨ I have VBN

Our Enriched Lemma Dictionary:

basturmaq = to crush [a rebellion]

Phrasal Translation Generation:

باستۇرغان ⇨ I have crushed [a rebellion]

(Hewitt, Post and Yarowsky, 2016)

Derivational Morphology

Derivational Morphology – Universalized Semantics

J:J(ATT) -ish
J:J(DIM) -ito
J:J(NEG) in-
J:J(NEG) un-
J:N(STATEQUALOF) -acity
J:N(STATEQUALOF) -ance
J:N(STATEQUALOF) -ancy
J:N(STATEQUALOF) -cy
J:N(STATEQUALOF) -dom
J:N(STATEQUALOF) -ence
J:N(STATEQUALOF) -ency
J:N(STATEQUALOF) -ern
J:N(STATEQUALOF) -ity
J:N(STATEQUALOF) -ness
J:N(STATEQUALOF) -ocity
J:N(STATEQUALOF) -sion
J:N(STATEQUALOF) -th
J:N(STATEQUALOF) -ty
J:R(INMANNER) -ily
J:R(INMANNER) -ly
J:V(CAUSETOBE) -ate
J:V(CAUSETOBE) -en
J:V(CAUSETOBE) -ify
J:V(CAUSETOBE) -ize
N:J(CHARBY) -some
N:J(FULLOF) -ful
N:J(FULLOF) -ious
N:J(FULLOF) -ous

N:J(HAVING) -ate
N:J(HAVING) -uous
N:J(LIKEA) -esque
N:J(LIKEA) -ish
N:J(LIKEA) -like
N:J(LIKEA) -oid
N:J(LIKEA) -ous
N:J(MADEOF) -y
N:J(QUALOF) -y
N:J(REALTEDTO) -ar
N:J(RELATEDTO) -al
N:J(RELATEDTO) -ual
N:J(RELATEDTO) -an
N:J(RELATEDTO) -ary
N:J(RELATEDTO) -ery
N:J(RELATEDTO) -ry
N:J(RELATEDTO) -ese
N:J(RELATEDTO) -etic
N:J(RELATEDTO) -atic
N:J(RELATEDTO) -ial
N:J(RELATEDTO) -ian
N:J(RELATEDTO) -ian
N:J(RELATEDTO) -ic
N:J(RELATEDTO) -ical
N:J(RELATEDTO) -ular
N:J(WITHOUT) -less
N:R(RELATEDTO) -ally
N:N(AUG-LARGE) mega-
N:N(AUG-SUPERIOR) over-

N:N(AUG-SUPERIOR) super-
N:N(DIM-INFERIOR) -ling
N:N(DIM-SMALL) -ette
N:N(DIM-SMALL) -ie
N:N(DIM-SMALL) -let
N:N(DIM-SMALL) -et
N:N(DIM-SMALL) -y
N:N(DOEROF) -ist
N:N(FEM) -ess
N:N(FEM) -ling
N:N(SMALLINSTANCEOF) -let
N:N(SMALLINSTANCEOF) -et
N:N(MATERIAL) -ing
N:N(REALMOF) -dom
N:N(ORIGIN) -ite
N:N(QUALITYOF) -ism
N:N(STATEQUALOF) -dom
N:N(STATEQUALOF) -hood
N:N(STATEQUALOF) -ship
N:N(WORKER-WITH) -man
N:N(WORKER-WITH) -boy
N:N(WORKER-WITH) -ier
N:N(WORKER-WITH) -eer
N:N(WORKER-WITH) -arian
N:N(RELATEDTO) -ory
N:R(INDIRECTIONOF) -ward
N:R(INDIRECTIONOF) -wise
N:V(CAUSETOHAVE) -ate
N:V(CAUSETOHAVE) -en
N:V(CAUSETOHAVE) -fy

Paradigms for Derivational Morphology

Concept	Lemma(V)	V:N(AGT)	V:N(PAT)	V:N(RES;ACTOF)	V:J(ABIL)
EMPLOY	employ	employer	employee	employment	employable
GIVE	give	giver	*recipient*	gift; giving	givable
TRANSPORT	transport	transporter	transportee	transportation	transportable
INTESTIGATE	investigate	investigator	investigated/N	investigation	investigable

Spanish: ↕

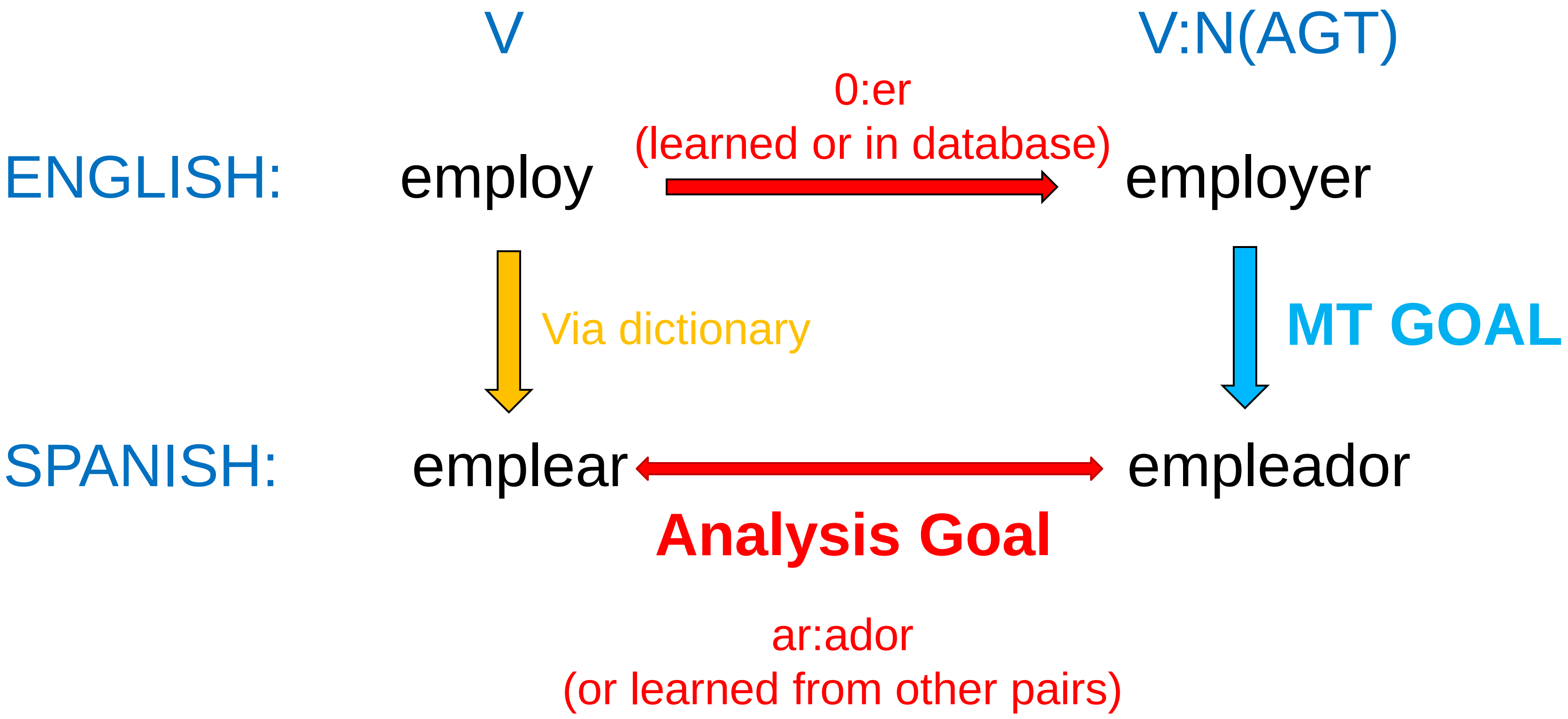
Concept	Lemma(V)	V:N(AGT)	V:N(PAT)	V:N(RES;ACTOF)	V:J(ABIL)
EMPLOY	emplear	empleador	empleado	empleo	empleable
GIVE	dar	dador	*receptor*	don;dar;*regalo*	dable
TRANSPORT	transportar	transportista	transportado	transporte	transportable
INTESTIGATE	investigar	investigador	investigado	investigación	investigable

Russian: ↕

Concept	Lemma(V)	V:N(AGT)	V:N(PAT)	V:N(RES;ACTOF)	V:J(ABIL)
EMPLOY	нанимать	наниматель	*работник*	*работа*	*трудоспособный*
GIVE	давать	даритель	данный	дарение	доступный
TRANSPORT	транспортировать	транспортер	транспортируемый	транспорт	транспортабельный
INTESTIGATE	исследовать	исследователь	исследуемый	исследование	…

Derivational Morphology

Learning Process:



V

V:N(AGT)

0:er

(learned or in database)

ENGLISH: employ → employer

Via dictionary

MT GOAL

SPANISH: emplear ↔ empleador

Analysis Goal

ar:ador

(or learned from other pairs)

Questions?